Genomic and phenotypic differentiation of the Aquilegia viridiflora complex along geographic distributions

Wei Zhang[1], Hua-Ying Wang[1], Tengjiao Zhang[1], Xiaoxue Fang[1], Meiying Liu[2], Mingzhou Sun[1], and Hongxing Xiao[1]

[1]Northeast Normal University
[2]Key Laboratory of Molecular Epigenetics of Ministry of Education, College of Life Sciences, Northeast Normal University

February 22, 2024

Abstract

How populations diverge into different lineages is a central issue in evolutionary biology. Despite the increasing evidence indicating that such divergences do not need geographic isolation, numerous phenotypic differentiations show a distributional correspondence. In addition, gene flow has been widely detected during and through such diverging processes. We used one widely distributed Aquilegia viridiflora complex as a model system to examine genomic differentiation and corresponding phenotypic variations along geographic gradients. Our phenotypic analyses of 90 individuals from 20 populations from northwest to northeast China identified two phenotypic groups along the geographic cline. All examined traits are distinct between them although a few intermediate individuals occur in their contacting regions. We further sequenced the genomes of the representative individuals of each population. However, we recovered four distinct genetic lineages based on both nuclear genomes and plastomes that were different from phenotypic differentiation. In particular, we recovered numerous genetic hybrids in the contact regions of four lineages. Gene flow is widespread and continuous between four lineages but much higher between contacting lineages than geographically isolated lineages. In addition, many genes with fast lineage-specific mutations were identified to be involved in local adaptation. Our results suggest that both geographic isolation and local selection exerted by the environment may together create geographic distributions of phenotypic variations as well as the underlying genomic divergences in numerous lineages.

Genomic and phenotypic differentiation of the *Aquilegia viridiflora*complex along geographic distributions

Wei Zhang [1], Huaying Wang [1], Tengjiao Zhang [1], Xiaoxue Fang [1]Meiying Liu [1], Mingzhou Sun [1] and Hongxing Xiao [1, *]

[1] Key Laboratory of Molecular Epigenetics of Ministry of Education, College of Life Sciences, Northeast Normal University, Changchun 130024, China

Correspondence: Hongxing Xiao, Key Laboratory of Molecular Epigenetics of Ministry of Education, College of Life Sciences, Northeast Normal University, Changchun 130024, China.

E-mail: xiaohx771@nenu.edu.cn

Running title: Lineage divergence of *A. viridiflora*complex

Abstract

How populations diverge into different lineages is a central issue in evolutionary biology. Despite the increasing evidence indicating that such divergences do not need geographic isolation, numerous phenotypic

differentiations show a distributional correspondence. In addition, gene flow has been widely detected during and through such diverging processes. We used one widely distributed *Aquilegia viridiflora*complex as a model system to examine genomic differentiation and corresponding phenotypic variations along geographic gradients. Our phenotypic analyses of 90 individuals from 20 populations from northwest to northeast China identified two phenotypic groups along the geographic cline. All examined traits are distinct between them although a few intermediate individuals occur in their contacting regions. We further sequenced the genomes of the representative individuals of each population. However, we recovered four distinct genetic lineages based on both nuclear genomes and plastomes that were different from phenotypic differentiation. In particular, we recovered numerous genetic hybrids in the contact regions of four lineages. Gene flow is widespread and continuous between four lineages but much higher between contacting lineages than geographically isolated lineages. In addition, many genes with fast lineage-specific mutations were identified to be involved in local adaptation. Our results suggest that both geographic isolation and local selection exerted by the environment may together create geographic distributions of phenotypic variations as well as the underlying genomic divergences in numerous lineages.

Key words: Geographic differentiation; phenotypes; genomes; gene flow; selection

INTRODUCTION

Disentangling the factors as drivers of genomic and phenotypic divergence is essential to understand speciation processes. Allopatric divergence has been considered the most likely cause of speciation for a long time, and targets of natural selection may contribute to allopatric speciation when populations encounter different selective pressures in different habitats (Wiens & Graham, 2005). One of the key speciation forces is evolutionary divergence driven by adaptation to the environment essentially proposed by Charles Darwin (1859). At the genomic level, loci with strong population differentiation reflect local adaptation in populations under different environments, in which reproductive isolation (RI) may appear as a byproduct of the accumulation of genetic differences (Schluter, 2001; Sobel, Chen, Watt, & Schemske, 2010; C. I. Wu, 2001). Natural selection thus acts as a "barrier" to gene flow to produce local genomic divergence between lineages (Arias, Van Belleghem, & McMillan, 2016; Edelman et al., 2019; Tavares et al., 2018). Similarly, gene flow and drift can also act and even interact in several ways during evolutionary divergence (Han et al., 2017; Ma et al., 2018). Geographic patterns of phenotype divergence also reflect that traits have diverged by environmental gradients as selective forces and created adaptive genetic differences (Ayoola et al., 2021; L.-F. Li, Cushman, He, & Li, 2020; T. Zhang et al., 2021). Extensive studies have indicated that morphological divergence is expected to be associated with prezygotic reproductive barriers. Plants can directly or indirectly reduce matting, sperm transfer, or fertilization (Coyne, 2016; Feder et al., 1994) by changing both the flowering time and mating system. For many taxa, including*Aquilegia* , floral characteristic differences leading to pollinator shifts likely contribute to RI between populations (Des Marais & Rausher, 2010; Hodges, Whittall, Fulton, & Yang, 2002; Kuriya, Hattori, Nagano, & Itino, 2015; Quattrocchio et al., 1999; Schwinn et al., 2006). If populations with divergent selection maintain geographic isolation, environmental differences may drive lineage adaptation and differentiation and play to form RI between them with restricted gene flow. Consequently, to make correct inferences about the driving force of lineage diversification, considering all processes along environmental gradients that shape phenotype and genome is important.

During population divergence, driving forces may issue distinct demographic histories, including divergence time, population size fluctuations, and directionality of gene flow. However, it can be difficult to identify which force is likely to have conduced to the current phylogeographical pattern of a certain species. The identification of highly differentiated regions in the genome may have false positive results due to the influence of the unique demographic histories (Krak et al., 2016). Therefore, it is important to reconstruct the past demography of lineages for plants in which interspecific gene flow has been detected widely. Moreover, East Asia occupies a diversiform climatic and geographical environment and is considered a natural laboratory for adaptive evolution (Ficetola, Mazel, & Thuiller, 2017). Evidence from previous studies in this region has uncovered the demographical histories, genetic diversities, and related influencing factors of many taxa (Areces-Berazain, Hinsinger, & Strijk, 2021; Song et al., 2021; S. Wu, Wang, Wang, Shrestha, & Liu, 2022).

While the heterogeneous environment of this area might have contributed to lineage divergence, it is still uncertain which might be the key driver.

Aquilegia viridiflora Pall. (Ranunculaceae), is a dominant, perennial herb that is widely distributed in northern China with obvious variation in phenotypes (Z. Wu, Raven, & Hong, 2001). In previous studies, Andrey S. Erst et al. identified *A. viridiflora* with purple laminae or lilac-blue petals as the new species *A. kamelinii* (A. Erst, Shaulo, & Schmakov, 2013) and *A. viridiflora* with dark purple petals in North China as the new species*A. hebeica* (A. S. Erst et al., 2017). However, the publication of these new species lacks support from molecular data. According to field surveys, the phenotypic variation of *A. viridiflora* is continuous (Figure S1) and presents substantial challenges in species delimitation. Therefore, *A. viridiflora* , *A. kamelinii* and *A. hebeica* were considered the species complex (*A. viridiflora*complex) in our study. In addition, *Aquilegia* is a well-known example of an evolutionary biology (Kramer, 2009). Phylogenetic studies have defined adaptive radiation in *Aquilegia* , involving a wide variety of habitats, and divergent selection played an important role in the radiation of *Aquilegia* (Fior et al., 2013; M. Li et al., 2019; W. Zhang, Wang, Dong, Zhang, & Xiao, 2021). Therefore, the*A. viridiflora* complex provides a remarkable system for assessing how evolutionary forces in various ways could have shaped genomic divergence patterns during speciation.

In the present study, we collected population samples covering the main distribution range of the *A. viridiflora* complex for genome resequencing and morphological characteristic measurements. Moreover, we assumed that environmental heterogeneity may put selective pressures on the *A. viridiflora* complex, driving population divergence and producing genetic variation. Thus, our study focuses on exploring phenotypic and genomic patterns of divergence, and especially on inferring the influence of gene flow, genetic drift, divergence time, divergent selection, and geographic isolation during the speciation process. The investigation reveals the contribution of evolutionary forces to the genomic divergence patterns of the *A. viridiflora*complex and helps us understand how demographics contribute to speciation.

MATERIALS AND METHODS

Study System

The *Aquilegia viridiflora* complex (Ranunculaceae), including*A. viridiflora* Pall., *A. hebeica* Erst and *A. kamelinii* Erst, is a perennial herbaceous plant throughout the mid-latitudes of the Northern Hemisphere. Here, to determine whether the*A. viridiflora* complex collected shared the most recent common ancestor (MRCA), we also collected other congener species whose distribution overlapped with that of the *A. viridiflora* complex, including*A. amurensis* Kom., *A. ecalcarata* Maxim., *A. japonica* Nakai et Hara, *A. oxysepala* Trautv. Et Mey. var. *kansuensis*Bruhl, *A. oxysepala* Trautv. Et Mey., and *A. yabeana* Kitag. seeds, as reported in our previous study (W. Zhang et al., 2021). Additionally, seeds of *Paraquilegia microphylla* were collected as an outgroup. Seeds of the *A. viridiflora* complex were collected from 20 locations covering its current distribution range, and all voucher specimens were identified by Dr Hongxing Xiao and deposited in the Northeast Normal University Herbaria (Figure 1A; Table S1). Each of the maternal plants was separated from the others by $>$ 50 m.

Common Garden Phenotypic Measurements and Statistics

All the seeds were sterilized and planted into pots under 12:12 light:dark conditions 25 °C/20 °C for 3 months. To obtain sufficient leaves and flowers, we transplanted the plants to outdoor agricultural fields at Northeast Normal University. For phenotypic measurement, 7 regenerative traits and 4 vegetative traits of different populations in the experimental field during the full flowering stage were measured. To ensure the accuracy of the measurements, we randomly selected 3-5 plants from each population to measure the number of inflorescences and plant height; among these plants, 6 flowers and 6 leaves were randomly selected from each plant, and the corolla diameter, pistil length, stamen length, leaf area, leaf perimeter and chlorophyll content were measured and recorded. Among these 6 flowers, we randomly selected 3 petals and calyxes to measure petal length (not including spurs), calyx length, spur length and calyx length.

To reduce the error caused by different measurement batches, we used a mixed linear model to evaluate traits

in the lme4 (Bates, Mächler, Bolker, & Walker, 2014) package in R according to the following regression model equation:

$$Y_i = \mu + \beta_1 + \beta_2 P + \varepsilon_i$$

whereY_i represents the traits of different populations, μ is the actual measurement, β_1 and β_2 are regression coefficients, A represents the measurement batches, P is the person conducting the measurement, and ε_i is the residual variance. The evaluation results were used for ANOVA and K-means cluster analysis in R.

DNA Sequencing and SNP Calling

Fresh leaves were collected from 1 population of *A. kamelinii* , 6 populations of *A. hebeica* and 13 populations of *A. viridiflora* to extract genomic DNA using a modified cetyltrimethylammonium bromide (CTAB) method (Doyle & Doyle, 1987), bringing the total number of sequence samples to 66. In addition, 12 individuals from other *Aquilegia* species overlapped the distribution of the *A. viridiflora* complex and 1 individual from*Paraquilegia microphylla* was also used to extract genomic DNA. For each individual, the Illumina Xten platform from Biomarker Technologies, Inc. (Beijing, China) was used for genomic library generation and sequencing with 2 × 150 bp paired reads. Furthermore, the raw sequence reads of*Semiaquilegia adoxoide* s (SRR437677) were downloaded from the NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra) to be used as an outgroup. To obtain high-quality genomes, all the reads were subjected to quality control by FastQC (Andrew, 2010) and filtered as follows: reads with adapters and reads with more than a 10% N content or more than 50% low-quality bases (quality value of less than 10) were removed. Low-quality reads were removed using NGStoolkit (Mulcare, 2004).

Clean sequence reads of 80 individuals were mapped to the reference genome of *A. coerulea* from the previous study of Filiault et al. (Filiault et al., 2018) using BWA v.0.7.12 with default parameters (H. Li & Durbin, 2009). SAMtools v.0.1.18 was used to convert SAM files to BAM files and sort reads (H. Li et al., 2009). The HaplotypeCaller, GenotypeGVCFs and CombineGVCFs modules in GATK v.4.1.8.0 were used to produce accurate SNP calls (McKenna et al., 2010). To improve the quality of SNPs, the VariantFiltration module in GATK v4.1.8.0 was used for filtration with the following parameters: "–filter-name FilterQual –filter-expression QUAL $<$ 30.0 –filter-name FilterQD –filter-expression QD $<$ 2.0 –filter-name FilterMQ –filter-expression MQ $<$ 40.0 –filter-name FilterFS –filter-expression FS $>$ 60.0 -window 5 -cluster 2". Next, VCFtools v0.1.13 (Danecek et al., 2011) was used to remove variants that 1) showed a minor allele frequency (MAF) of 0.02 or less, 2) were not balletic variants, 3) showed a sequencing depth of less than 5, and 4) showed a missing rate exceeding 0.5.

Population Genetic Structure Analysis

Individuals named by species and location, we estimated the phylogenetic relationships of other*Aquilegia* species with the *A. viridiflora* complex to determine whether the *A. viridiflora* complex in our study shared an MRCA using IQ-TREE multicore version 1.6.12 (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) and MEGA X (Kumar, Stecher, Li, Knyaz, & Tamura, 2018) with 1000 bootstrap replicates. Both the ML tree and NJ tree indicated that 20 populations of the *A. viridiflora* complex shared an MRCA. All trees were illustrated in iTOL (http:// itol.embl.de). Therefore, 672,439 high-quality SNPs in 20 populations of the *A. viridiflora* complex were used for downstream analysis after removing other *Aquilegia* species. To explore the patterns of genetic structure of the *A. viridiflora* complex, we used a phylogenetic network by the Neighbor-Net algorithm in the software Splits Tree (Huson & Bryant, 2006) with 1000 bootstrap replicates. ADMIXTURE v.1.3.0 (Alexander, Novembre, & Lange, 2009) was applied to investigate the maximum likelihood of the ancestry of all individuals with K values ranging from 2 to 10 with 10 replicates for each K value and examined the optimum K value according to the lowest value of the error rate. Principal component analysis (PCA) was performed using EIGENSOFT v.6.1.4 (Price et al., 2006) to infer population genetic structure. By combining the results of the phylogenetic relationship and genetic structure analysis to establish lineages for downstream analysis, we divided 20 populations into four lineages. Since the results of the population genetic structure showed a mixed genetic background in some individuals, the Python

package HyDe was used to identify hybridization events at the individual level (Blischak, Chifman, Wolfe, & Kubatko, 2018). Among them, P2 was an individual who did not show a mixed genetic background, and P1 and P3 were individuals of other populations, respectively.

Phylogeny of *A. viridiflora* complex chloroplast genomes

Taking the chloroplast genome of *A. viridiflora* (MN809220) as a reference genome, clean reads were aligned to the reference to obtain the variation of *A. viridiflora* complex in the chloroplast genome and called SNPs according to the above pipeline. SNPs in the chloroplast genome were also used to infer the phylogeny of the *A. viridiflora* complex by Splits Tree, and the chloroplast genomes of*Paraquilegia microphylla* and *Semiaquilegia adoxoides* were regarded as outgroups (Huson & Bryant, 2006). Haplotype network analysis was performed using Median Joining Network in PopART v 1.7 to analyze the *A. viridiflora* complex with default parameters (Leigh & Bryant, 2015).

Demographic History Inference

Hybrid individuals were removed from downstream analyses due to the levels of heterozygosity, and the remaining individuals were used for the inference of demographic history. The gene flow between each lineage was inferred based on the allele frequency by applying the ABBA-BABA or D statistic using the qpDstat module in AdmixTools 7.0 (Patterson et al., 2012). Based on the phylogeny (((P1, P2), P3), O), the four topologies were tested, with *Paraquilegia microphylla* and*Semiaquilegia adoxoides* as outgroups. In addition to the D statistic, we performed gene flow analysis between different populations using allele frequency data with 1-10 migration events by TreeMix v1.1 (Pickrell & Pritchard, 2012). The Python script easySFS was used to calculate the joint site frequency spectrum (SFS) for demographic analysis (https://github.com/isaacovercast/easySFS). We also calculated the likelihood of different demographic scenarios in fastsimcoal2 software (Excoffier et al., 2021) using the joint SFS to infer demographic parameters. Twenty scenarios were set up involving genetic structure and gene flow, including three monophyletic models and sixteen paraphyletic models (Figure S8). For each scenario, fastsimcoal2 performed 10000 coalescent simulations to approximate the expected SFS in each cycle and will run 40 optimization cycles to estimate the parameters. To ensure the accuracy of evaluating the best scenario, each scenario was run 100 times, and the run with the highest likelihood was compared by calculating the Akaike information criterion (AIC) to determine the best scenario. The parameter estimation was run under the best scenario 100 times with each of bootstrapped SFS. A neutral mutation rate of 10^{-8} and a generation time of 1 year were used to estimate the effective population size, divergence times and migration rates (M. Li et al., 2019).

Adaptation Analysis

Nucleotide diversity (π) was calculated for the four groups using a 100 kb nonoverlapping sliding window by using VCFtools (Danecek et al., 2011). Additionally, the fixation index (F_{ST}) and nucleotide divergence (D_{XY}) between each of the four groups were calculated by PIXY (Danecek et al., 2011) in 100 kb sliding windows. Ninety-five percent confidence intervals for mean F_{ST} and D_{XY} values were obtained by bpnreg packages in R. PopLDdecay (C. Zhang, Dong, Xu, He, & Yang, 2019) software was applied to compute linkage disequilibrium (LD) decay among different groups and chromosomes separately. To make the following analyses easier to complete, we employed PLINK v.1.9 (Purcell et al., 2007) to filter SNPs with the following parameters: –indep-pairwise 50 10 0.2, 143,318 SNPs were kept and phased using Beagle v.3.3.2 (Browning & Browning, 2007). We ran Mantel tests to test for correlations between F_{ST} and geographic distance to assess isolation by distance (IBD) in the R package vegan (Oksanen et al., 2013). The Q_{ST} were calculate using R package '$Q_{ST}F_{ST}$Comp' (Gilbert & Whitlock, 2015).

To explore the effect of adaptation on 11 quantitative traits, we used the single-phenotype Q_{ST}-F_{ST}test with the R package '$Q_{ST}F_{ST}$Comp' (Gilbert & Whitlock, 2015). If $Q_{ST} > F_{ST}$, the differentiation of traits is the major effect of divergent selection and shows local adaptation. We used the half-sib dam model and 10000 resampling steps for each Q_{ST}-F_{ST} analysis. Taking into account the recent divergence between NE and EL lineages, we compared the levels of F_{ST} between CN and NW lineages to identify candidate loci under natural selection from 143,318 SNPs by BayeScan v.2.1 (Foll & Gaggiotti, 2008) software with default

parameters, and PGDSpider (Lischer & Excoffier, 2012) was used to produce an input file for BayeScan. SNPs with a q value lower than 0.05 were considered potentially selected loci. Fisher's exact test was used to perform significance in GO enrichment analysis on positively selected genes by the clusterProfiler package (Yu, Wang, Han, & He, 2012) in R.

Environmental Association Analysis

Temperature and precipitation are considered the driving factors that affect the population growth rate and limit the distribution of species (Cahill et al., 2014; Dalgleish, Koons, Hooten, Moffet, & Adler, 2011; Kim & Donohue, 2013). Therefore, nineteen current bioclimatic variables with a spatial resolution of 30 s were collected from WorldClim (*http://www* .worldclim.org/).Simultaneously, we recorded the GPS information of the sampling locations and downloaded the GPS information of *A. viridiflora* from CVH (*http://www*.cvh.ac.cn) and GBIF (*http://www*.gbif.org). ArcMap v.10.4 was used to limit the spatial extent according to the buffer radius (5 km) around each occurrence record. We used $|r| < 0.8$ (Pearson correlation coefficient) as a cutoff to remove highly correlated variables. The seven retained current bioclimatic factors (Bio1: annual mean temperature; Bio2: mean diurnal range; Bio3: isothermality; Bio4: temperature seasonality; Bio8: mean temperature of the wettest quarter; Bio15: precipitation seasonality; Bio17: precipitation of the driest quarter) were used for subsequent analysis. Redundancy analysis (RDA) was used to assess the impact of current bioclimatic factors on the genomic composition of the *A. viridiflora* complex in R. For the same reason as selection analysis, we also used CN and NW lineages to identify the loci related to the seven retained current bioclimatic factors. BAYENV2 (Coop, Witonsky, Di Rienzo, & Pritchard, 2010) was used with 1,000,000 iterations and run three times separately. For each bioclimatic factor, the SNPs among the top 1% according to BF and among the top 5% according to the absolute Spearman's ρ were considered candidates. Candidate SNPs were mapped to the corresponding genes, and GO enrichment was performed by the clusterProfiler package (Yu et al., 2012) in R.

RESULTS

Phenotypic Variation across the Distribution

We sampled twenty populations of the *A. viridiflora* complex with various phenotypes that covered its entire distribution (Figure 1A and Table S1). The phenotypes of *A. viridiflora* complex populations grown in common gardens were investigated and compared, and each phenotype differed significantly among different populations by using ANOVA (Table 1). K-means cluster analysis was used to visualize the phenotypic variation among populations and identified two distinctive groups. Dim1 and Dim2 could explain 76.6% and 19.5% of the observed variation, respectively, and the cumulative contribution to the observed phenotypic variation was 96.1%. All individuals of *Aquilegia viridiflora* and *A. kamelinii* were found in Cluster 1, while individuals of *A. hebeica* were found in Cluster 2 (Figure 1A and S2A). There was a significantly negative correlation coefficient between some floral characteristics, including corolla length, petal length, spur length and pistil length, and the nutritional traits, including leaf area, leaf perimeter and height. In addition, the number of inflorescences was significantly negatively correlated with the above floral characteristics, while it was significantly positively correlated with the plant height (Figure S3).

Four lineages can be identified in the *A. viridiflora*complex

Individuals from twenty populations of the *A. viridiflora* complex were sampled for resequencing. Additionally, to ensure that the species complex with various phenotypes could be regarded as a monophyletic group, we also sampled other sympatric wild columbine species with the*A. viridiflora* complex. Whole-genome sequencing of 80 individuals from nine species was performed, and after filtering, we obtained 1,064,089 high-quality biallelic single nucleotide polymorphisms (SNPs). We constructed the phylogenetic relationships among the *Aquilegia*species through the ML and NJ methods based on nuclear genome SNPs. Both topologies indicated that *A. viridiflora* , *A. kamelinii* and*A. hebeica* shared an MRCA with strong support and relationships among species were close to each other (Figure S4). Therefore, 672,439 high-quality biallelic SNPs in the 20 populations of the *A. viridiflora* complex were used to assess their evolution. Functional annotations indicated that 55.604% of SNPs were located upstream and downstream, while 17.454% were

in intronic regions, and 7.07% were in exonic regions of genes. The ML tree inferred from the above SNPS indicated that the individuals of the *A. viridiflora* complex were divided into four lineages (NE, EL, CN and NW): NE comprised *A. kamelinii* and the individuals of *A. viridiflora* distribution in northeastern China, EL comprised the individuals of *A. viridiflora* and *A. hebeica* distribution in East Shandong South Liaoning area, the individuals of *A. hebeica* distribution in North China belonged to a single lineage (EL), and the individuals of*A. viridiflora* distribution in northwestern China belonged to a single lineage (NW). In this case, the *A. viridiflora* complex showed a paraphyletic pattern, that is, NE and EL formed a sister clade, and the other two lineages, CN and NW, were closely related (Figure 1C). This revealed a different evolutionary history from the clusters based on phenotypes.

The population genetic structure of the *A. viridiflora* complex indicated that ancestral clustering at $K = 4$ was optimal according to the cross-validation error rate (Figure S5). The result of ancestral inference was obviously consistent with the geographical distribution of the 20 populations and the phylogenetic relationships detailed above. Individuals of SZ, LT, YM, XW, HL, HH and HD populations in the contact regions of lineages showed multiple ancestral compositions (Figure 1A and 1B), which might reflect recent gene flow between these lineages. From the PCA plot, the first principal component (PC1) and second principal component (PC2) explained 6.61% and 4.04% of the observed variation, respectively. It also showed four distinct lineages among the 20 populations, while individuals with multiple ancestral compositions occupied an intermediate space in distinct lineages (Figure S2B). The neighbor-net phylogenetic network depicted these patterns by grouping the differential of NE and EL, while CN and NW were not clearly differentiated, at the same time, it was also proven that the above individuals had a mixed genetic background (Figure 2A). In addition, we also detected a significant signal of hybridization in the above populations at the individual level, in which P1 and P2 did not belong to a group with hybrids (Table S2). Unsurprisingly, the neighbor-net phylogenetic network based on 190 polymorphic sites in the chloroplast genome also showed a little difference from that based on genome polymorphisms resulting from hybridization and backcrossing of hybrid lineages but still showed a paraphyletic pattern (Figure S6A). Taken together, there are four lineages across the sampled *A. viridiflora* complex through chloroplast and genome polymorphisms.

To understand the diversity patterns, the nucleotide diversity (π) of the NE, EL, CN and NW lineages was calculated throughout the genome for each 50 kb with a 10 kb step-size. Among the four lineages, lineage NW showed the highest nucleotide diversity, and lineage EL showed the lowest nucleotide polymorphism (Figure 2B). Based on chloroplast genome polymorphisms, we detected 27 haplotypes among the 66 *A. viridiflora* complex individuals. The haplotype diversity (hd) and nucleotide diversity (π) for all individuals were 0.971 and 0.00022, respectively. Among the four lineages, NW had the highest haplotype diversity and nucleotide diversity (Table S3). Moreover, the haplotypes in NW were in this network elsewhere, while the haplotypes in the other groups were limited in the haplotype network (Figure S6B, Table S4).

Tests of gene flow and dynamic history inference

To characterize the demographic histories of the four lineages, we removed hybrids to investigate gene flow, effective population size (Ne) and divergence time. We first examined the gene flow between lineages through the ABBA-BABA test. For four lineages, (((NE, EL), CN), outgroup), (((CN, NW), NE), outgroup) and (((CN, NW), EL), outgroup) showed a significant deviation of D-stat from zero (absolute value of Z-score greater than 3), indicating that gene flow exists between lineages EL and CN, CN and NE, CN and EL. Interestingly, no gene flow was detected between the NW lineage and other lineages (Figure 3A and Table S5). Next, we employed TreeMix to further investigate the complex patterns of gene flow between populations. The results of TreeMix analysis inferred that recent gene flow was only exhibited in the EL and NE lineages, while historical gene flow was exhibited in NE, EL and CN lineages (Figure S7).

Combined with the results of the ABBA-BABA test and TreeMix analysis, the evolutionary history of the four lineages was inferred by fastsimicoal2 using pairwise joint site frequency spectra. Based on the best-supported model (Table S6), the effective population size was 927,589 for the ancestors. The ancestral population was differentiated into eastern and western lineages at approximately 239 Kya (95% highest posterior density (HPD) = 217–267 Kya), and their effective population size were 385,321 and 433,093, respectively. Next,

CN and NW separated at approximately 211 Kya (95% HPD = 196–227 Kya), and NE and EL separated at 168 Kya (95% HPD = 153–184 Kya). All the divergence times were in the Middle Pleistocene. The current effective population sizes of NE, EL, CN and NW were 561,674, 99,440, 433,094 and 385,321, respectively (Figure 3B and Table S7). In addition, the divergence of the four lineages accompanied six bidirectional gene flow events inferred by fastsimcoal2, including three ancient gene flow and three modern gene flow. The modern gene flow was higher from NE to EL than others (CN to EL > CN to NE > EL to NE > NE to CN > EL to CN), which indicated gene flow between contacting lineages than those geographically isolated ones. The credible lineage divergence pattern had a better fit based on the comparison of the simulated dataset with the observed site frequency spectra (Figure S9).

F_{ST} was calculated across different lineages (NE, EL, CN and NW) to infer population genetic differentiation. At the overall level of the genome, the F_{ST} values had the highest value between EL and NW (> CN vs. NW > NE vs. EL > EL vs. CN > NE vs. NW > NE vs. CN), which indicated that the *A. viridiflora* complex had a moderate level of differentiation. Lineage NW was the most differentiated from the other lineages may be due to the lack of recent gene flow. This pattern was also obvious from the $F_{ST}$and D_{XY} values calculated for the four lineages using 10 kb windows across genomes, which were consistent with F_{ST} at the overall level of the genome (Figure 3C). We compared linkage disequilibrium decay between different chromosomes of four lineages. The analysis showed that CN and EL presented a greater degree of LD, while NE and NW showed less linkage disequilibrium (indicated by r^2). When $r^2 = 0.1$, the decay distances of NE, EL, CN and NW were 10 kb, 37 kb, 44 kb and 6.9 kb, respectively (Figure 3D). The rapid decay of NE and NW may have been due to the higher genetic diversity relative to that of EL and CN.

Local phenotypic adaptation

Common garden experiments were conducted to investigate the phenotypic divergence evidence for genome differentiation and adaptation of the*A. viridiflora* complex. The mean F_{ST} for the four lineages was 0.1519 (95% HPD: 0.14949-0.15074), and the overall Q_{ST} of the six traits was higher than the mean F_{ST}, including the corolla diameter, petal length, spur length, pistil length, inflorescence number and leaf area, indicating that local adaptation was driven by these traits (Table 1). The highest Q_{ST} was procured by measuring the corolla diameter, followed by the spur length, petal length, pistil length, inflorescence number and leaf area, which also showed a higher Q_{ST} of floral characteristics than nutritional traits. Among the four lineages, CN showed the smallest corolla diameter and the shortest petal length, spur length and pistil length, followed by EL, while NW and NE showed the largest values of these phenotypes (the difference between NW and NE was not significant) (Figure S10A - D). Interestingly, the number of inflorescences showed the reverse order of the above traits (Figure S10E). EL showed the largest leaf area, followed by CN, while the difference in leaf area between NW and NE was not significant (Figure S10F).

Genes under the selection and driven by the environment

Despite being closely related, CN and NW differ in morphology and ecology (Figure S10). To explore the genetic mechanism of the differentiation, we identified 1168 outlier SNPs from the 143,318 filtered SNPs according to a 0.05 threshold for the q-value between the CN and NW lineages based on the Bayesian method applied in BAYESCAN (Figure 4A). The outlier SNPs were in 487 genes that were identified under positive selection (Table S8). The F_{ST} values of outlier SNPs were significantly higher than those of others, which suggests that the divergence of the lineages from their ancestral population was possibly caused by strong divergent selection. GO enrichment analyses of positively selected genes were significantly enriched in ABC-2 type transporter family protein (GO:0016887, p= 0.0151), and we then integrated the candidate genes with organ development, reproductive isolation, biotic and abiotic stress responses. Among these genes , several (e.g., *PAD2* (Parisy et al., 2007), *DRB3* (Mehdi, Reza, Hassan, Shima, & Gholamreza, 2022) and *EDM2* (Eulgem et al., 2007)) are involved in plant immunity, whereas some are associated with the stress of drought (e.g.,*MAPKKK21* (L. Wu, Chang, Wang, Wu, & Wang, 2021) and *SCL1*(Manohara Reddy et al., 2018)) and cold (e.g., *CER3* (Rahman et al., 2021) and *LOS1* (Bielsa, Ávila-Alonso, Fernández i Martí, Grimplet, & Rubio-Cabetas, 2021)). Moreover, two genes (*GRF2*(Beltramino et al., 2018) and *PHT4;2* (Irigoyen, Karlsson, Kuruvilla, Spetea, & Versaw, 2011)) participate in the regulation of leaf size,

and several candidate genes are associated with flowering time (e.g., *FT* (Tyagi et al., 2018), *FLK* (Lim et al., 2004), *MYB30* (L. Zhu et al., 2020) and *FY* (Henderson, Liu, Drea, Simpson, & Dean, 2005)), flower organ size(*KNU* (Bollier et al., 2018) and *CKX5* (Bartrina, Otto, Strnad, Werner, & Schmülling, 2011)), flower color (*MYB113* (Jiao, Zhao, Wu, Song, & Li, 2020)) and pollen development (*WLIM1* (Yang et al., 2022) and *EXO70A2* (Marković et al., 2020)). The rapid divergence of candidate genes with reproductive function might drive prezygotic isolation across lineages.

Additionally, the Mantel test revealed a significant effect of IBD on the genome-wide SNPs of the *A. viridiflora* complex (Figure 4B). Because genetic divergence can result from selection driven by heterogeneous environments via geographic distance, redundancy analysis (RDA) was then implemented to examine the bioclimatic factors related to the 143,318 SNPs. The contribution of seven environmental variables in RDA space is shown in Figure S11. The strongest predictor was the mean diurnal range (Bio2), followed by isothermality (Bio3), mean temperature of the wettest quarter (Bio8), temperature seasonality (Bio4), precipitation seasonality (Bio15), annual mean temperatures (Bio1) and precipitation of the driest quarter (Bio17). BAYENV2 identified 179 SNPs significantly associated with the 7 examined environmental variables (Figure 4C). The outlier SNPs were in 83 genes and significantly enriched in "PEBP (phosphatidylethanolamine-binding protein) family protein" (GO:0003712, 2 genes, $p = 0.034$), which was previously known to be involved in various physiological processes, such as seasonal growth (Khosa, Bellinazzo, Kamenetsky Goldstein, Macknight, & Immink, 2021), seed germination (B. Zhang, Li, Li, & Yu, 2020) and floral transition (Y. Zhu, Klasfeld, & Wagner, 2021). In addition, seven genes were adaptively differentiated under temperature and water stress, indicating that the crosstalk between the two modules observed might be important in the process of local adaptation (Table S9).

DISCUSSION

Our population genetic analyses of the *A. viridiflora* complex within its main range across Northern China suggest a paraphyletic divergence scenario of the species complex with significant phenotypic divergence. These lineages show moderate levels of genetic differentiation and diverged from each other 168 to 239 Kya, a timing in the Middle Pleistocene that the most important climate transition has occurred, namely, the "mid-Brunhes transition" (MBT, ˜400 Kya) (Ao et al., 2020). The results further indicate that both geographic isolation and selective pressures via environmental factors have contributed to the observed lineage differentiation. In the following, these findings will be discussed in more detail.

Lineage divergence across geographic distribution

Four lineages were identified with phylogenetic, Bayesian clustering approaches and PCA based on nuclear SNPs and chloroplast genomes: the NE lineage included *A. viridiflora* and *A. kamelinii* , the EL lineage included *A. viridiflora* and *A. hebeica* , and the CN and NW lineages only included *A. hebeica* and *A. viridiflra* , respectively, which is different from taxonomic recognition species (Figure 1). The moderate genetic divergence (both F_{ST} and D_{XY}) among different lineages may be explained by two nonmutually exclusive factors, isolation-by-distance and gene flow. While a strong pattern of isolation-by-distance (IBD) was observed across all of the study populations and could explain the variance in genetic differentiation (Figure 4B), gene flow after lineage differentiation may be decrease the level of genetic differentiation. We took advantage of the whole-genome dataset of the *A. viridiflora*complex to infer the amount of gene flow of the recovered lineages, revealing that a higher level of recent gene flow than ancient was shown to have occurred. Hybrids with two or three lineages mixed in the contact zone existed because of the absence of strong biogeographical barriers. Chloroplast capture might result in the chloroplast genome of populations in the contact zone exhibiting inconsistent clustering with nuclear SNPs, which also supported recent hybridization occurring in the contact zone. In particular, gene flow between pairs of lineages with close geographic distribution ranges was relatively large, indicating that geographic isolation might shape the present range and phylogeny of the recovered lineages. Although no recent gene flow was detected between NW and the other three lineages (Figure 3A-B), hybrids which formed by NE, CN and NW lineage mixing still arose, which resulted from the ancient gene flow not only between the ancestral clade of NE and EL lineages and NW lineage, but also between the ancestral clade of NE and EL lineages and the ancestral clade

of CN and NW lineages. Additionally, the level of divergence between the CN and NW lineages was higher than that between the other two lineages according to the F_{ST}value. D_{XY} was also evaluated the divergence between pairs of lineages and was consistent with the F_{ST} value (Figure 3C). Therefore, the incipient lineage divergence of the *A. viridiflora* complex could be largely explained by geographic isolation and the decrease of gene flow caused by the increase of divergence time. These results are in line with recent studies suggesting that gene flow may have occurred between currently geographically isolated East Asia and played a major role in other plant radiations (Hu et al.; Xiao et al., 2020; L. L. Xu et al., 2021).

Local selection and adaptation

The *A. viridiflora* complex was already present in Asia by 1.19 Mya is evident from the age model of Fior et al. in accordance with molecular clock estimates of the flora of Altai (*A. glandulosa* ,*A. sibirica* , and *A. viridiflora*) (Fior et al., 2013). It is clear that present lineages must have diverged much more recently through our demographic history simulations. The model indicates that a very young divergence time in the *A. viridiflora* complex: c. 239 Kya, c. 211 Kya, and c. 168 Kya (Figure 3B). Although these ages must be explained with caution, the most important climate transition occurred in the Middle Pleistocene, which might have resulted in lineage divergence by changing the suitable habitat of species. The role of Middle Pleistocene climate transitions in speciation processes has long been advised, such as the *Ranunculus auricomus* complex (Tomasello, Karbstein, Hodač, Paetzold, & Hörandl, 2020), *Populus rotundifolia* (J. L. Li et al., 2021) and *Cerapanorpa brevicornis*(Gao, Hua, Xing, & Hua, 2022). Here, we add to these "species on the speciation way" examples in the recent Middle Pleistocene speciation and indicate that the very high rates of speciation associated with *Aquilegia* adaptive radiation might be driven by the Middle Pleistocene climate transition. Based on our observation, the CN lineage representing *A. hebeica* and the NW lineage representing *A. viridiflora* , which have experienced heterogeneous environments and different environmental variables, exert differential selection pressure on the different lineages. We found that 83 genes related to environmental factors that may play a key role in the continuously adaptive process (Figure 4C). Some genes associated with the abscisic acid (ABA) signaling pathway (*AIN1* (Dong et al., 2021) and *AAO4* (Seo et al., 2004)) can regulate numerous ABA responses and may induce the abiotic stress responses for defense in different environments.

Apart from the genetic divergence revealed by phylogenetic and population structure analyses, clear differentiation in phenotypic traits was also exhibited based on common garden data (Figure 1). Divergence in corolla diameter, petal length, spur length, pistil length, inflorescence number and leaf area was probably driven by selection. Among these traits, floral characteristic differentiation might act as a prezygotic isolation mechanism between the four lineages. Additionally, we found 487 genes under positive selection and exhibiting high divergence between lineages (Figure 4A). These genes were significantly enriched on ABC-2 type transport family protein, a gene family involved in a wide range of metabolism in plants and playing import roles in seed germination, stomatal movement, lateral root development and responses to various environmental stresses (Liu, Li, & Liu, 2013; Matsuda et al., 2012; X. D. Zhang, Zhao, & Yang, 2018). Among these selected genes, *KNU* (Bollier et al., 2018) and*CKX5* (Bartrina et al., 2011) are involved in the regulation of flower morphology and development, of which a major difference exists between the CN and NW lineages. Similarly, *GRF2* (Beltramino et al., 2018) and *PHT4;2* (Irigoyen et al., 2011) initiate fixation in those lineages due to their relation to leaf development and morphology. *NRPD1B* is related to panicle branches (L. Xu et al., 2020), and high allelic divergence and fixation of this gene in CN lineages may contribute to producing more inflorescence relative to NW lineages. Moreover, several genes (e.g., *PAD2* (Parisy et al., 2007), *DRB3* (Mehdi et al., 2022) and *EDM2* (Eulgem et al., 2007)) are involved in disease resistance in response to biotic stresses. Plants can adjust growth and defense based on different environments to survive and reproduce in the natural world, which might contribute to lineage divergence (He, Webster, & He, 2022). Therefore, our study illuminated that geographic isolation and local selection drove the lineage divergence of the *A. viridiflora* complex and created geographic distributions of phenotypic variations. Further work is needed to acquire more accurate functions of genes in the whole genome under the selective pressure and clarify potential adaptation patterns.

Conclusions

In summary, we explored the phenotypic and genetic diversification of 20 populations of the *A. viridiflora* complex across its range. Our results revealed lineage divergence in the *A. viridiflora*complex, and four lineages were distinguished with a high geographic correlation and most likely derived from divergent selection due to heterogeneous environments and geographic isolation. In addition, continuous gene flow between lineages existed during the lineage divergence. By conducting multiple analyses, we identified the genetic variables associated with environmental variables, which might influence the evolution process. More importantly, our findings provide genomic insights into the integration of information on gene flow, divergence time of lineages, population effect size, and divergent selection to identify genes potentially involved in speciation and the basis of geographic distributions of phenotypic variations. This study may also accelerate further taxonomic and evolutionary studies on the *A. viridiflora* complex and other species in the East Asia.

AUTHOR CONTRIBUTIONS

X. HX. and W. HY designed the study and evaluated the results; S. MZ, W. HY and Z. W. collected the materials; Z. W., Z. TJ., F. XX. And L. MY participated in data analysis; Z. W. and W. HY. prepared the manuscript; all authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We are grateful to Jianquan Liu for suggestion and comments to improve the manuscript; Shu Wang for collecting the plant materials; and Peng Peng for providing photographs. The research was supported by the National Natural Science Foundation of China (32070244) and "the Fundamental Research Funds for the Central Universities".

DATA AVAILABILITY STATEMENT

Raw sequence data is available from the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) under the submission PRJNA756091. The phenotypic and climate data have been archived in dryad (https://doi.org/10.5061/dryad.sqv9s4n4m).

ORCID

Hongxing Xiao https://orcid.org/0000-0002-6040-5443

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research, 19* (9), 1655-1664.

Andrew, S. (2010). A quality control tool for high throughput sequence data. *Fast QC, 532* .

Ao, H., Rohling, E. J., Stringer, C., Roberts, A. P., Dekkers, M. J., Dupont-Nivet, G., . . . Liu, Z. (2020). Two-stage mid-Brunhes climate transition and mid-Pleistocene human diversification.*Earth-Science Reviews, 210* , 103354.

Areces-Berazain, F., Hinsinger, D. D., & Strijk, J. S. (2021). Genome-wide supermatrix analyses of maples (Acer, Sapindaceae) reveal recurring inter-continental migration, mass extinction, and rapid lineage divergence. *Genomics, 113* (2), 681-692.

Arias, C. F., Van Belleghem, S., & McMillan, W. O. (2016). Genomics at the evolving species boundary. *Current opinion in insect science, 13* , 7-15.

Ayoola, A. O., Zhang, B.-L., Meisel, R. P., Nneji, L. M., Shao, Y., Morenikeji, O. B., . . . Okeyoyin, A. O. (2021). Population genomics reveals incipient speciation, introgression, and adaptation in the African Mona Monkey (Cercopithecus mona). *Molecular biology and evolution, 38* (3), 876-890.

Bartrina, I., Otto, E., Strnad, M., Werner, T., & Schmülling, T. (2011). Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana. *The Plant Cell, 23* (1), 69-80.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* .

Beltramino, M., Ercoli, M. F., Debernardi, J. M., Goldy, C., Rojas, A. M., Nota, F., . . . Palatnik, J. F. (2018). Robust increase of leaf size by Arabidopsis thaliana GRF3-like transcription factors under different growth conditions. *Scientific Reports, 8* (1), 1-13.

Bielsa, B., Ávila-Alonso, J. I., Fernández i Martí, Á., Grimplet, J., & Rubio-Cabetas, M. J. (2021). Gene Expression Analysis in Cold Stress Conditions Reveals BBX20 and CLO as Potential Biomarkers for Cold Tolerance in Almond. *Horticulturae, 7* (12), 527.

Blischak, P. D., Chifman, J., Wolfe, A. D., & Kubatko, L. S. (2018). HyDe: a Python package for genome-scale hybridization detection.*Systematic biology, 67* (5), 821-829.

Bollier, N., Sicard, A., Leblond, J., Latrasse, D., Gonzalez, N., Gévaudant, F., . . . Chevalier, C. (2018). At-MINI ZINC FINGER2 and Sl-INHIBITOR OF MERISTEM ACTIVITY, a conserved missing link in the regulation of floral meristem termination in Arabidopsis and tomato.*The Plant Cell, 30* (1), 83-100.

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American journal of human genetics, 81* (5), 1084-1097.

Cahill, A. E., Aiello-Lammens, M. E., Caitlin Fisher-Reid, M., Hua, X., Karanewsky, C. J., Ryu, H. Y., . . . Wiens, J. J. (2014). Causes of warm-edge range limits: systematic review, proximate factors and implications for climate change. *Journal of Biogeography, 41* (3), 429-442.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation.*Genetics, 185* (4), 1411-1423.

Coyne, J. A. (2016). Theodosius Dobzhansky on hybrid sterility and speciation. *Genetics, 202* (1), 5-7.

Dalgleish, H. J., Koons, D. N., Hooten, M. B., Moffet, C. A., & Adler, P. B. (2011). Climate influences the demography of three dominant sagebrush steppe plants. *Ecology, 92* (1), 75-85.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158.

Des Marais, D. L., & Rausher, M. D. (2010). Parallel evolution at multiple levels in the origin of hummingbird pollinated flowers in Ipomoea. *Evolution: International Journal of Organic Evolution, 64* (7), 2044-2054.

Dong, T., Yin, X., Wang, H., Lu, P., Liu, X., Gong, C., & Wu, Y. (2021). ABA-INDUCED expression 1 is involved in ABA-inhibited primary root elongation via modulating ROS homeostasis in Arabidopsis.*Plant Science, 304* , 110821.

Doyle, J. J., & Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue* . Retrieved from

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., . . . Neafsey, D. E. (2019). Genomic architecture and introgression shape a butterfly radiation. *science, 366* (6465), 594-599.

Erst, A., Shaulo, D., & Schmakov, A. (2013). *Aquilegia kamelinii*(Ranunculaceae)–a new species from North Asia. *Turczaninowia, 16* (3), 8-10.

Erst, A. S., Wang, W., Yu, S. X., Xiang, K., Wang, J., Shaulo, D. N., . . . Nobis, M. (2017). Two new species and four new records of*Aquilegia* (Ranunculaceae) from China. *Phytotaxa, 316* (2), 121-137.

Eulgem, T., Tsuchiya, T., Wang, X. J., Beasley, B., Cuzick, A., Tor, M., . . . Dangl, J. L. (2007). EDM2 is required for RPP7-dependent disease resistance in Arabidopsis and affects RPP7 transcript levels. *The Plant Journal, 49* (5), 829-839.

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics, 37* (24), 4882-4885.

Feder, J. L., Opp, S. B., Wlazlo, B., Reynolds, K., Go, W., & Spisak, S. (1994). Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. *Proceedings of the National Academy of Sciences, 91* (17), 7990-7994.

Ficetola, G. F., Mazel, F., & Thuiller, W. (2017). Global determinants of zoogeographical boundaries. *Nature Ecology & Evolution, 1* (4), 1-7.

Filiault, D. L., Ballerini, E. S., Mandakova, T., Akoz, G., Derieg, N. J., Schmutz, J., . . . Hayes, R. D. (2018). The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife, 7* , e36426.

Fior, S., Li, M., Oxelman, B., Viola, R., Hodges, S. A., Ometto, L., & Varotto, C. (2013). Spatiotemporal reconstruction of the*Aquilegia* rapid radiation through next-generation sequencing of rapidly evolving cp DNA regions. *New Phytologist, 198* (2), 579-592.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics, 180* (2), 977-993.

Gao, K., Hua, Y., Xing, L. X., & Hua, B. Z. (2022). Speciation of the cold-adapted scorpionfly Cerapanorpa brevicornis (Mecoptera: Panorpidae) via interglacial refugia. *Insect Conservation and Diversity, 15* (1), 114-127.

Gilbert, K. J., & Whitlock, M. C. (2015). QST–FST comparisons with unbalanced half-sib designs. *Molecular ecology resources, 15* (2), 262-267.

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome research, 27* (6), 1004-1015.

He, Z., Webster, S., & He, S. Y. (2022). Growth–defense trade-offs in plants. *Current Biology, 32* (12), R634-R639.

Henderson, I. R., Liu, F., Drea, S., Simpson, G. G., & Dean, C. (2005). An allelic series reveals essential roles for FY in plant development in addition to flowering-time control.

Hodges, S. A., Whittall, J. B., Fulton, M., & Yang, J. Y. (2002). Genetics of floral traits influencing reproductive isolation between Aquilegia formosa and Aquilegia pubescens. *the american naturalist, 159* (S3), S51-S60.

Hu, H., Yang, Y., Li, A., Zheng, Z., Zhang, J., & Liu, J. Genomic divergence of Stellera chamaejasme through local selection across the Qinghai-Tibet Plateau and northern China. *Molecular ecology* .

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution, 23* (2), 254-267.

Irigoyen, S., Karlsson, P. M., Kuruvilla, J., Spetea, C., & Versaw, W. K. (2011). The sink-specific plastidic phosphate transporter PHT4; 2 influences starch accumulation and leaf size in Arabidopsis. *Plant physiology, 157* (4), 1765-1777.

Jiao, F., Zhao, L., Wu, X., Song, Z., & Li, Y. (2020). Metabolome and transcriptome analyses of the molecular mechanisms of flower color mutation in tobacco. *BMC Genomics, 21* (1), 1-10.

Khosa, J., Bellinazzo, F., Kamenetsky Goldstein, R., Macknight, R., & Immink, R. G. (2021). PHOSPHATIDYLETHANOLAMINE-BINDING PROTEINS: the conductors of dual reproduction in plants with vegetative storage organs. *Journal of experimental botany, 72* (8), 2845-2856.

Kim, E., & Donohue, K. (2013). Local adaptation and plasticity of*Erysimum capitatum* to altitude: its implications for responses to climate change. *Journal of Ecology, 101* (3), 796-805.

Krak, K., Vit, P., Belyayev, A., Douda, J., Hreusova, L., & Mandak, B. (2016). Allopolyploid origin of *Chenopodium album* s. str.(Chenopodiaceae): a molecular and cytogenetic insight. *PLoS One, 11* (8), e0161063.

Kramer, E. M. (2009). *Aquilegia* : a new model for plant development, ecology, and evolution. *Annual review of plant biology, 60* , 261-277.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms.*Molecular biology and evolution, 35* (6), 1547-1549.

Kuriya, S., Hattori, M., Nagano, Y., & Itino, T. (2015). Altitudinal flower size variation correlates with local pollinator size in a bumblebee-pollinated herb, Prunella vulgaris L.(Lamiaceae).*Journal of evolutionary biology, 28* (10), 1761-1769.

Leigh, J. W., & Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6* (9), 1110-1116.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25* (14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools.*Bioinformatics, 25* (16), 2078-2079.

Li, J. L., Zhong, L. L., Wang, J., Ma, T., Mao, K. S., & Zhang, L. (2021). Genomic insights into speciation history and local adaptation of an alpine aspen in the Qinghai–Tibet Plateau and adjacent highlands.*Journal of Systematics and Evolution, 59* (6), 1220-1231.

Li, L.-F., Cushman, S. A., He, Y.-X., & Li, Y. (2020). Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping forsythia. *Horticulture research, 7* .

Li, M., Wang, H., Ding, N., Lu, T., Huang, Y., Xiao, H., . . . Li, L. (2019). Rapid divergence followed by adaptation to contrasting ecological niches of two closely related columbine species*Aquilegia japonica* and *A. oxysepala* . *Genome biology and evolution, 11* (3), 919-930.

Lim, M.-H., Kim, J., Kim, Y.-S., Chung, K.-S., Seo, Y.-H., Lee, I., . . . Park, C.-M. (2004). A new Arabidopsis gene, FLK, encodes an RNA binding protein with K homology motifs and regulates flowering time via FLOWERING LOCUS C. *The Plant Cell, 16* (3), 731-740.

Lischer, H. E., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics, 28* (2), 298-299.

Liu, S., Li, Q., & Liu, Z. (2013). Genome-wide identification, characterization and phylogenetic analysis of 50 catfish ATP-binding cassette (ABC) transporter genes. *PLoS One, 8* (5), e63895.

Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., . . . Abbott, R. J. (2018). Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex.*Proceedings of the National Academy of Sciences, 115* (2), E236-E243.

Manohara Reddy, B., Kumari, V., Anthony Johnson, A., Jagadeesh Kumar, N., Venkatesh, B., Jayamma, N., & Sudhakar, C. (2018). Scarecrow like Protein 1,(Ct-SCL1) Involved in Drought Stress Tolerance by Interacting with SWI3B Component of Chromatin Modelling Complex in Cluster Bean, Cyamopsistetragonaloba (L.) Taub. *Int. J. Res. Anal. Rev, 5* .

Marković, V., Cvrčková, F., Potocký, M., Kulich, I., Pejchar, P., Kollárová, E., . . . Žárský, V. (2020). EXO70A2 is critical for exocyst complex function in pollen development. *Plant physiology, 184* (4), 1823-1839.

Matsuda, S., Funabiki, A., Furukawa, K., Komori, N., Koike, M., Tokuji, Y., . . . Kato, K. (2012). Genome-wide analysis and expression profiling of half-size ABC protein subgroup G in rice in response to abiotic stress and phytohormone treatments. *Molecular genetics and genomics, 287* (10), 819-835.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research, 20* (9), 1297-1303.

Mehdi, R. M., Reza, Y. A., Hassan, M. M., Shima, M., & Gholamreza, K. (2022). Disease Prevention, Genetic Selection, and Vaccination Based on BoLA-DRB3. 2 Polymorphism: A Model for Immunogenetic Studies.

Mulcare, D. M. (2004). NGS Toolkit, Part 8: The National Geodetic Survey NADCON Tool. *Professional Surveyor Magazine* .

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution, 32* (1), 268-274.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., . . . Wagner, H. (2013). Package 'vegan'. *Community ecology package, version, 2* (9), 1-295.

Parisy, V., Poinssot, B., Owsianowski, L., Buchala, A., Glazebrook, J., & Mauch, F. (2007). Identification of PAD2 as a γ-glutamylcysteine synthetase highlights the importance of glutathione in disease resistance of Arabidopsis. *The Plant Journal, 49* (1), 159-172.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics, 192* (3), 1065-1093.

Pickrell, J., & Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings* , 1-1.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics, 38* (8), 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics, 81* (3), 559-575.

Quattrocchio, F., Wing, J., van der Woude, K., Souer, E., de Vetten, N., Mol, J., & Koes, R. (1999). Molecular analysis of the anthocyanin2 gene of petunia and its role in the evolution of flower color. *The Plant Cell, 11* (8), 1433-1444.

Rahman, T., Shao, M., Pahari, S., Venglat, P., Soolanayakanahally, R., Qiu, X., . . . Tanino, K. (2021). Dissecting the roles of cuticular wax in plant resistance to shoot dehydration and low-temperature stress in Arabidopsis. *International Journal of Molecular Sciences, 22* (4), 1554.

Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution, 16* (7), 372-380.

Schwinn, K., Venail, J., Shang, Y., Mackay, S., Alm, V., Butelli, E., . . . Martin, C. (2006). A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus Antirrhinum. *The Plant Cell, 18* (4), 831-851.

Seo, M., Aoki, H., Koiwai, H., Kamiya, Y., Nambara, E., & Koshiba, T. (2004). Comparative studies on the Arabidopsis aldehyde oxidase (AAO) gene family revealed a major role of AAO3 in ABA biosynthesis

in seeds.*Plant and cell physiology, 45* (11), 1694-1703.

Sobel, J. M., Chen, G. F., Watt, L. R., & Schemske, D. W. (2010). The biology of speciation. *Evolution: International Journal of Organic Evolution, 64* (2), 295-315.

Song, X., Milne, R. I., Fan, X., Xie, S., Zhang, L., Zheng, H., . . . Ma, T. (2021). Blow to the Northeast? Intraspecific differentiation of Populus davidiana suggests a north-eastward skew of a phylogeographic break in East Asia. *Journal of Biogeography, 48* (1), 187-201.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., . . . Li, M. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences, 115* (43), 11006-11011.

Tomasello, S., Karbstein, K., Hodač, L., Paetzold, C., & Hörandl, E. (2020). Phylogenomics unravels Quaternary vicariance and allopatric speciation patterns in temperate-montane plant species: a case study on the Ranunculus auricomus species complex. *Molecular ecology, 29* (11), 2031-2049.

Tyagi, S., Mazumdar, P. A., Mayee, P., Shivaraj, S., Anand, S., Singh, A., . . . Kumar, A. (2018). Natural variation in Brassica FT homeologs influences multiple agronomic traits including flowering time, silique shape, oil profile, stomatal morphology and plant height in B. juncea.*Plant Science, 277* , 251-266.

Wiens, J. J., & Graham, C. H. (2005). Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual review of ecology, evolution, and systematics* , 519-539.

Wu, C. I. (2001). The genic view of the process of speciation.*Journal of evolutionary biology, 14* (6), 851-865.

Wu, L., Chang, Y., Wang, L., Wu, J., & Wang, S. (2021). Genetic dissection of drought resistance based on root traits at the bud stage in common bean. *Theoretical and applied genetics, 134* (4), 1047-1061.

Wu, S., Wang, Y., Wang, Z., Shrestha, N., & Liu, J. (2022). Species divergence with gene flow and hybrid speciation on the Qinghai–Tibet Plateau. *New Phytologist, 234* (2), 392-404.

Wu, Z., Raven, P., & Hong, D. (2001). *Flora of China. Volume 6* : Science Press.

Xiao, J. H., Ding, X., Li, L., Ma, H., Ci, X. Q., van der Merwe, M., . . . Li, J. (2020). Miocene diversification of a golden-thread nanmu tree species (Phoebe zhennan, Lauraceae) around the Sichuan Basin shaped by the East Asian monsoon. *Ecology and Evolution, 10* (19), 10543-10557.

Xu, L., Yuan, K., Yuan, M., Meng, X., Chen, M., Wu, J., . . . Qi, Y. (2020). Regulation of rice tillering by RNA-directed DNA methylation at miniature inverted-repeat transposable elements. *Molecular plant, 13* (6), 851-863.

Xu, L. L., Yu, R. M., Lin, X. R., Zhang, B. W., Li, N., Lin, K., . . . Bai, W. N. (2021). Different rates of pollen and seed gene flow cause branch-length and geographic cytonuclear discordance within Asian butternuts. *New Phytologist, 232* (1), 388-403.

Yang, X., Bu, Y., Niu, F., Cun, Y., Zhang, L., & Song, X. (2022). Comprehensive analysis of LIM gene family in wheat reveals the involvement of TaLIM2 in pollen development. *Plant Science, 314* , 111101.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters.*Omics: a journal of integrative biology, 16* (5), 284-287.

Zhang, B., Li, C., Li, Y., & Yu, H. (2020). Mobile TERMINAL FLOWER1 determines seed size in Arabidopsis. *Nature plants, 6* (9), 1146-1157.

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., & Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics, 35* (10), 1786-1788.

Zhang, T., Chen, J., Zhang, J., Guo, Y. T., Zhou, X., Li, M. W., . . . Nevo, E. (2021). Phenotypic and genomic adaptations to the extremely high elevation in plateau zokor (Myospalax baileyi). *Molecular ecology, 30* (22), 5765-5779.

Zhang, W., Wang, H., Dong, J., Zhang, T., & Xiao, H. (2021). Comparative chloroplast genomes and phylogenetic analysis of Aquilegia.*Applications in plant sciences, 9* (3), e11412.

Zhang, X. D., Zhao, K. X., & Yang, Z. M. (2018). Identification of genomic ATP binding cassette (ABC) transporter genes and Cd-responsive ABCs in Brassica napus. *Gene, 664* , 139-151.

Zhu, L., Guan, Y., Liu, Y., Zhang, Z., Jaffar, M. A., Song, A., . . . Chen, F. (2020). Regulation of flowering time in chrysanthemum by the R2R3 MYB transcription factor CmMYB2 is associated with changes in gibberellin metabolism. *Horticulture research, 7* .

Zhu, Y., Klasfeld, S., & Wagner, D. (2021). Molecular regulation of plant developmental transitions and plant architecture via PEPB family proteins: an update on mechanism of action. *Journal of experimental botany, 72* (7), 2301-2311.

FIGURE LEGENDS

Figure 1. Sampling localities and population genetic structure of the *Aquilegia viridiflora*complex. A. Geographical distributions of the sampled *A. viridiflora* complex, where colors in circles distinguish groups. The genetic structure showed ADMIXTURE proportions of genetic clusters for each individual of the four lineages at the best K value (K = 4), and each bar represents an individual. The scale in the figure is 1:20,000,000, B. Population genetic structure inferred by ADMIXTURE when K= 3-5, C. phylogenetic relationships of *A. viridiflora* complex according to the maximum likelihood (ML) method, and blue, red, light green and orange represent the NE, EL, CN and NW lineages, respectively. The asterisk indicates that the bootstrap value is 100.

Figure 2. A. NeighborNet diagram based on genome, B. nucleotide diversity (ϑ_π) of*A. viridiflora* complex.

Figure 3. Demographic history of *Aquilegia viridiflora* complex and population genetic analysis. A. Results for the D-statistics. The solid red dot represents that the absolute value of Z-scores is greater than 3, and the hollow red dot represents that the absolute value of Z-scores is less than 3, B. ancestral population sizes, population divergence times, and migration rates were assessed by fastimcoal2, C. pairwise F_{ST} and D_{XY} between different groups, D. LD decay of the four lineages of *A. viridiflora* complex. Except that the difference between the two groups marked in the figure is not significant, there is a significant difference between any other two groups.

Figure 4. Genomic regions with signals of selection and adaptation to local environmental conditions. A. Bayescan results for genome-wide scans to detect outlier SNPs based on locus-specific FST values. Red dots represent qval less than 0.05 and indicated the SNPs under the positive selection, B. relationship of genetic distance and geographical distance, C. outlier SNPs associated with bioclimate factors. The red dots indicate that the SNP in the sliding window was selected by the environment; gray dots represent other regions in the genome. The map shows the spatial variation in different bioclimate factors.

Table legends:

Table 1. The results of Qst-Fst analysis.

Supplemental information:

Figure S1 Phenotypes of the different populations of *Aquilegia viridiflora* complex in the common garden.

Figure S2 A. k-means cluster analysis of phenotype based on the first two principal components, B. Principal component analysis (PCA) plot for the 66 *A. viridiflora* complex individuals based on the first two principal

components.

Figure S3 Phylogenetic relationship between *Aquilegia viridiflora* complex and other columbine species whose distribution was overlap with *Aquilegia viridiflora* complex. A. Neighbor-joining (NJ) tree, B. maximum likelihood (ML) tree.

Figure S4 Correlation analysis between phenotypes. * represents a significant relationship between different phenotypes, $0.01 < p < 0.05$. ** represents a significant relationship between different phenotypes, $0.001 < p < 0.01$. *** represents a significant relationship between different phenotypes, $p < 0.001$.

FigureS5 Cross-validation results corresponding to different K values in the ADMIXTURE analysis of *A. viridiflora* complex.

Figure S6 A. Median-joining haplotype network of 66 *A. viridiflora* complex based on chloroplast genome, B. NeighborNet diagram based on chloroplast genome.

Figure S7 Five gene flow events of *Aquilegia viridiflora* complex. Each branch represents a population, and arrows indicate migration events that occur between populations.

Figure S8 A. Four groups of possible models for inferring demographic parameters. The gray arrows represent the current gene flow, the red arrows represent the ancient gene flow, and the model in the dotted line is the best model selected according to the values of the likelihoods and AIC.

Figure S9 Comparison of observed data and expected data based on the best model for A. NE and EL, B. NE and CN, C. EL and CN, D. NE and NW, E. EL and NW, and F. CN and NW.

Figure S10 Phenotype of *Aquilegia viridiflora* complex. A. Corolla diameter, B petal length (not including spur), C. angle between petals and spurs, D. spur length, E. pistil length, F. Number of inflorescences. Except that the difference between the two groups marked in the figure is not significant, there is a significant difference between any other two groups.

Figure S11 RDA plot of *A. viridiflora* complex showing SNPs with are potentially associated with seven environmental factors.

Table S1 Sampling information and the value of seven environmental factors.

Table S2. Results of the HyDe.

Table S3. Polymorphisms in the chloroplast genome of *A. viridiflora* complex.

Table S4. Haplotype in the chloroplast genome of *A. viridiflora* complex.

Table S5. Results of the D-statistics.

Table S6. The AIC value of different models shown in Figure S7.

Table S7. Inferred parameters estimates with 95% confidence intervals for the best-fitting demographic scenario modelled in fastasimcoal2.

Table S8. The gene list under the selection.

Table S9. The gene list correlation the environment.

Hosted file

`Table.docx` available at https://authorea.com/users/500580/articles/581322-genomic-and-phenotypic-differentiation-of-the-aquilegia-viridiflora-complex-along-geographic-distributions